Conditional Switching: A New Variety of Regression with Many Potential Environmental Applications

Michael E. Tarter,[1] Michael D. Lock,[2] and Rose M. Ray[3]

[1]Department of Biomedical and Environmental Health Sciences, University of California, Berkeley, CA 94720 USA; [2]Becton Dickinson Immunocytometry Systems, San Jose, CA 95131 USA; [3]Failure Analysis Associates, Menlo Park, CA 94025 USA

We introduce a new form of regression that has many applications to environmental studies. For a sequence composed of key variates with prototypic value x, this form differs from the estimation of a location parameter-based curve, $\mu(x)$, a scale parameter-based curve, $\sigma(x)$, or other currently used types of regression. Instead of estimating a curve location, scale, or α–quantile parameter, it assumes that there are two or more population subgroups; for example, consisting of unsensitized and sensitized individuals, respectively. Although within each subgroup the relationships $\mu(x)$ or $\sigma(x)$ may or may not be horizontal, these relationships are not deemed to be of primary importance. Instead, the mixing parameter P that indexes the proportions of the two subgroups is treated as being related to the key variate value x.

In the sense that its goal is the estimation of a proportion, the new procedure resembles logit regression. But, in terms of the continuous spectrum of values attained by the response variate, the means used to attain its goal are dissimilar from those of logit regression. Specifically, group membership is not known directly but is determined from a proxy continuous variate whose values overlap between groups. Examples are given with simulated and natural data where this new form of regression is applied. We believe that conditional switching regression is a particularly valuable research tool when chemical level x of an induced asthma attack or birthweight x measured in a study of the biomarker cotinine's effect on pregnancy outcomes determines whether an attack or a negative outcome occurs. In these applications it is this binary-valued and yet indirectly measured response, and not the continuously valued attack severity or birthweight deficit variate, that is of primary interest. *Key words*: birth weight, Fourier series, gestation age, mixture decomposition, nonparametric estimation, nonparametric regression, population components, switching regression. *Environ Health Perspect* 103:748–755 (1995)

The research advances that are described in this paper can be said to constitute a border between modern statistical points of view and recent developments in digital computation. The latter had not occurred when Tarter et al. (*1*) described one common pitfall that occurs in environmental research when skew-shaped frequency curves are encountered. It is also often the case that the data configuration dealt with here involves skew-shaped frequency curves.

For example, consider the estimated density of a gestation age variate, measured in days, shown in Figure 1. The exact nature of the data and the methodology used to study these data will be described below. One particular feature of this figure is worth noting at this point. Rather than the smoothly decreasing right tail of the lognormal density, the estimated curve appears to have a bump or shoulder on its right side. As discussed by Tarter (*2*), the curve resembles the cross-section of a rug that has been laid over two distinct distributional components.

Suppose that a mother's recall of gestation age is not best described by a single, albeit skewed, density model like the lognormal. Suppose instead that the data obtained from some mothers are accurate, while data obtained from their counterparts are a collection of guesses. Within both the accurate and guess subgroups there will be considerable variation. Even if they were typically overestimates, there will be some guesses that are smaller in value than accurate reports, and there will be some accurate reports of gestation age that are larger in magnitude than some guesses. Consequently, the problem here is not simply an outlier, censoring, or truncation issue.

Now consider the realistic situation where it is not merely a univariate curve, such as that shown in Figure 1, that is of interest. Suppose there is a dose level, exposure level, or key variate such as birthweight that is associated with the response variate like gestation age. The same mothers who tend to guess rather than accurately report gestation age may also tend to give birth to underweight babies. Were this to be the case, as will be argued below, then this statistical configuration has many obvious, and several subtle, effects on research studies.

Methods

Suppose the conditional probability density of a response variate value $Y = y$ at a given value x of dose, or key variate, X, is represented by

$$f(y|x) = P(x)\frac{f_1\{[y-\mu_1(x)]/\sigma_1(x)\}}{\sigma_1(x)} + [1-P(x)]\frac{f_2\{[y-\mu_2(x)]/\sigma_2(x)\}}{\sigma_2(x)}, \quad (1)$$

where f_1 and f_2 are probability densities that are symmetric about zero. Here regression functions $\mu_1(x)$, $\sigma_1(x)$, $\mu_2(x)$, and $\sigma_2(x)$ describe the subgroup-specific parametric variation with value x of variate X. On the other hand, the mixing parameter regression function $P(x)$ expresses the relationship between, 1) the Y variate's attribution to density $f_1\{[y - \mu_1(x)]/\sigma_1(x)\}/\sigma_1(x)$ or density $f_2\{[y - \mu_2(x)]/\sigma_2(x)\}/\sigma_2(x)$, and, second, the value of the key variable X. Thus, if the indexing variate values, $i = 1,2$; characterize the pair of densities $f_i\{[y - \mu_i(x)]/\sigma_i(x)\}/\sigma_i(x)$, $i = 1,2$, then the curve $P(x)$ discloses the probability that an individual selected at random from the subpopulation whose X variate equals x will prove to be a member of the $i = 1$ group.

The problem dealt with here is more complex than that dealt with by logit regression at least in two respects. For example, consider the situation where mothers specify the time between their children's conception and birth. When a response of y days is obtained, the value is not accompanied by the response: ". . . my answer is a guess." Only the value y itself is available.

Also, unlike logit regression estimation of the curve $\log\{P(x)/[1 - P(x)]\}$, the highly complex representation of $f(y|x)$ in Equation 1, and not the more tractable representation of $\log\{P(x)/[1 - P(x)]\}$ by an elementary function, is utilized. Thus, besides the curve $P(x)$, the four curves $\mu_1(x)$, $\sigma_1(x)$, $\mu_2(x)$, and $\sigma_2(x)$ form part of the model that also contains the function $P(x)$. The major assertion of this paper is that modern statistical and computational methodology can now be used to tease the structure of the curve $P(x)$ apart from its counterpart curves $\mu_1(x)$, $\sigma_1(x)$, $\mu_2(x)$, and $\sigma_2(x)$.

Clearly, no method can successfully

Address correspondence to R.M. Ray, Failure Analysis Associates, PO Box 3015, Menlo Park, CA 94025 USA.

This research was supported by National Institute of Environmental Health Sciences grant 1 RO1 ES05379. We thank Brenda Eskenazi for helping us gain access to the data.

Received 25 August 1994; accepted 27 April 1995.

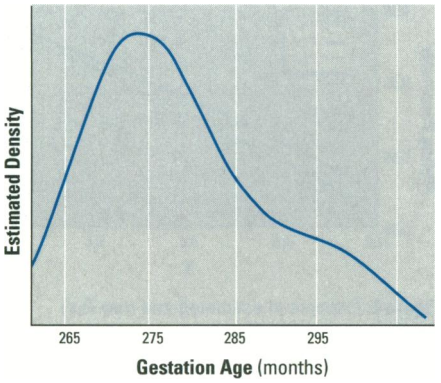

Figure 1. Gestation age mixture density

estimate $P(x)$ for every possible configuration of $f_1(x)$, $f_2(x)$. Yet the following method can be used in a large number of applications. This technique has a long history, but it has been applied infrequently in environmental science.

Background

Buchanan-Wollaston and Hodgeson (*3*) used the estimated distribution modes to decompose mixtures of normals. Eisenberger (*4*) used both mode and trough location (a point at which an estimated curve attains a local minimum) for this purpose. Weichselberger (*5*) used graphs of the ratios between estimated density evaluations. Weichselberger's method uses ratios of sample frequencies from adjacent equal-length class intervals to approximate a straight line plotted on semilogarithmic paper.

The assumption that the tail-region overlap of mixture components is inappreciable has been used by Harding (*6*), Preston (*7*), and Cassie (*8*) to devise graphical techniques for estimating the leftmost or rightmost component location parameters. The idea that numerical or graphical subtraction procedures could be used to sequentially deal with mixture decomposition was first discussed by Szigeti (*9*) and Labhart (*10*), while Noble et al. (*11*) compared estimated densities by overlaying cathode ray tube displays with curves printed on transparent paper.

As far as nongraphical approaches are concerned, Cicchinelli (*12*) and Hasselblad (*13*) describe likelihood procedures, and Frankowski (*14*) constructed moment-based procedures for mixture decomposition based in part on an in-depth survey of previous methods.

Between Frankowski's work and the advent of PC-based computation, considerable work had been done on the problem of mixture decomposition. Informative reviews and applications of this work appear in Titterington et al. (*15*). It is noteworthy, however, that much of this recent research concerning decomposition focuses on the problem of multimodality. As shown by Tukey (*16*) and commented on by Tarter and Lock (*17*), it is common to encounter multimodal curves that are not actually mixtures, just as it is common for mixtures to occur that are not bimodal or multimodal.

Although his work did not directly deal with either data analysis or computational considerations, Medgyessy (*18*) brought the lambda method (the approach emphasized in this paper), which was first proposed by the mathematician Doetsch (*19,20*), to the attention of the applied statistical community. However, Doetsch's papers, Medgyessy's 1961 book, and a subsequent paper by Gregor (*21*) made no distinction between the problem of decomposing a mixture where the actual function could be directly evaluated at an equally spaced sequence of points and the problem of mixture decomposition where the data are obtained from an independent, identically distributed sample.

The problem of applying Doetsch's Fourier-based method to sample data was first considered by Kronmal (*22*). Two papers based on Kronmal's approach were published by Stanat (*23,24*). The advantages of applying this technique in the bivariate case were discussed by Tarter and Silvers (*25*), and Titterington et al. (*15*); Tarter et al. (*26*) and Tarter (*27*) introduced all but one of the specific techniques described below.

Mixture Decomposition, Cluster Analysis, and Regression

The primary difference between the approaches outlined in the previous section and those considered here is best seen in the perspective of cluster analysis. The popular procedure of cluster analysis can be viewed as a heuristic mixture decomposition technique that has several appealing features when applied to high-dimensional data sets. Yet it is based on the presumption that any given data point should be associated with one particular cluster. In comparison to the all-or-nothing inclusion presumed by clustering approaches, the curve $P(x)$ can be viewed as a "fuzziness index," in the sense that the word "fuzzy" is defined as "not clear in shape, especially at the edges" (*28*). In the two-component case, $P(x) = 1$ and $P(x) = 0$, respectively, denote certainty of inclusion or noninclusion in the first group. What is new here is that the fuzziness index $P(x)$ is treated as a regression curve.

The practicality of $P(x)$ curve estimation depends on a feature of the form of curve representation first selected by Doetsch (*19,20*). Although Fourier coefficients of conditional densities can be expressed very simply and conveniently in terms of their joint distributional counterparts, Tarter (*29*) suggested that this may have been overlooked. By incorporating the Fourier coefficients, one can obtain rather conveniently a sequence of estimated conditionals (i.e., slices) of a high-dimensional density. In practical terms, each slice shown in this paper was obtained in a response time of less than 1 sec on an IBM PS/2 Model 70 with a 16 MHz processor. On more recently available Intel 66 MHz 486DX2 and Pentium 60 MHz processors, our routines can generate a sequence of 50 slices in less than 1 sec.

In addition to the capacity to execute programs that implement conditional switching regression on very economical systems, running time is nearly independent of the sample sizes involved. This useful feature is due to the step that is intermediate to the actual slicing process. What is sliced is not the data set itself, but a Fourier or frequency domain representation of the data set. As is elaborated upon in Tarter and Lock (*17*), the number of sample Fourier coefficients actually used in all computations is much smaller than the actual data sets from which these coefficients are estimated.

As explained in Appendix B, for $c > 1$, the lambda-based conditional switching regression's applicability to the c-component mixture model

$$f\left(y|p_1,p_2,\ldots,p_{c-1},\mu_1,\mu_2,\ldots,\mu_c,\sigma_1,\sigma_2,\ldots,\sigma_c\right) = p_1 f_1\left[\left(y-\mu_1\right)/\sigma_1\right]/\sigma_1 + p_2 f_2\left[\left(y-\mu_2\right)/\sigma_2\right]/\sigma_2 + \ldots + \left(1-\sum_{j=1}^{c-1} p_j\right) f_c\left[\left(y-\mu_c\right)/\sigma_c\right]/\sigma_c \quad (2)$$

depends on two classes of assumptions. The first is that unimodal f_1, f_2,... are symmetric about a mode with coordinate 0. The second is that the parameters σ_1^2, σ_2^2,..., σ_c^2, are individually proportional or have some other monotonic relationship to the population variances (second cumulants) of the respective subcomponents.

Under these assumptions, the parameters σ_1^2, σ_2^2,..., σ_c^2 of the mixture can be altered by specific lambda user control settings in a manner that reduces subcomponent overlap. Even though the value of c may not actually be known prior to use of the lambda-based conditional switching regression, the structure of the mixture model (specifically, the addition operation that separates mixture components) ensures that a single multiplier sequence will distribute among c-components. Like a prescribed drug that is taken orally or injected into the bloodstream, the lambda-based conditional switching regression is designed to target precisely those portions of individual mixture components, the

parameters σ_1^2, σ_2^2,..., σ_c^2, which, when reduced, expedite mixture decomposition. Like most drugs, this process does not work satisfactorily in all situations. For example, it cannot directly deal with mixtures of components that have identical location parameters but different scale parameters because in these situations, even though the common value of all component location parameters can be changed, this change will not help separate mixture components.

Theory

The methodological approach used to estimate the function $P(x)$ is based on the distinction between mixing parameter and other forms of regression. For example, in the bivariate case, what Quandt (*30,31*) refers to as switching parameter regression involves the decomposition of a mixture of two bivariate normal densities, where the constant value $P(x) = p$ specifies how much one density contributes to the overall mixture. Specifically, in Quandt and Ramsey (*32*), the mixing parameter p is not considered to be functionally related to the value of the random variate. The decomposition of bivariate normals is discussed by Tarter and Silvers (*25*) and Titterington et al. (*15*). The two-variate special case for which $P(x) = p$ and $P(x)$ is therefore a horizontal line is the only situation where mixing parameter regression is equivalent to the decomposition of bivariate normal densities.

A mixing parameter regression curve $\hat{P}(x)$ and the curve it estimates $P(x)$, the normal $N(\mu = 0, \sigma = 1.5)$ cumulative distribution function $P(x) \equiv \Phi(x/1.5)$, are both shown in Figure 2. While the x coordinates of the curves shown in this figure are similar to the x coordinates of other forms of regression, the y coordinates correspond to the true values of $P(x)$, the dashed curve, and the estimated values, $\hat{P}(x)$, the solid curve.

Before turning to the details of the algorithm by which these figures were obtained, a sequence of examples using simulated as opposed to natural data will be informative. To illustrate the procedure used to generate the solid curve shown in Figure 2, it is useful to show how the Quandt form of switching regression differs from the form discussed in this paper. The Quandt approach assumes that mixing parameter P is not functionally related to values, x, taken on by the independent variate X, while the estimation of $\hat{P}(x)$ is not based on this assumption.

In all the simulated data examples the within-component correlation was set equal to 0. The sample size of $n = 400$ was selected to match the sample size used in the natural data trials described below.

The parameter settings used to obtain configurations shown in both Figures 3 and 4 resulted in mixtures of bivariate data, where for Figure 3 the probability of assignment to the mixture component whose y mean equaled 0.5, irrespective of the value of x, was the same identical value, specifically $P(x) \equiv 0.5$. In both Figures 3 and 4 the x-variate component means were both set equal to 0 and the x-variate component variances were set equal to 1. However, in order to clearly demonstrate the difference between the Quandt and new forms of switching regression, the y-component standard deviations were both set equal to the uncharacteristically small value, 0.1. One feature of the simulation routine that was used to obtain the data set displayed in Figure 4 that differs from its Figure 3 counterpart is that the curve $P(x)$ corresponding to Figure 4 switches from y = 0 to y = 1 within a small neighborhood of the x coordinate $x = 0$. Consequently in the Figure 4 case there is a very small chance that a data point will be attributable to the upper component when its x coordinate is less than zero while, on the other hand, points such that $x > 0$ are, in

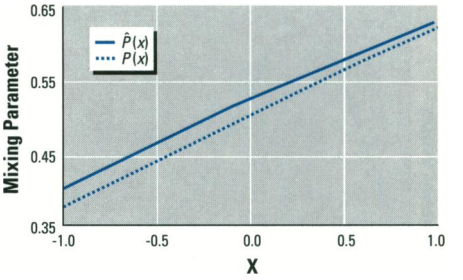

Figure 2. Example of estimated and true $P(x)$ curves.

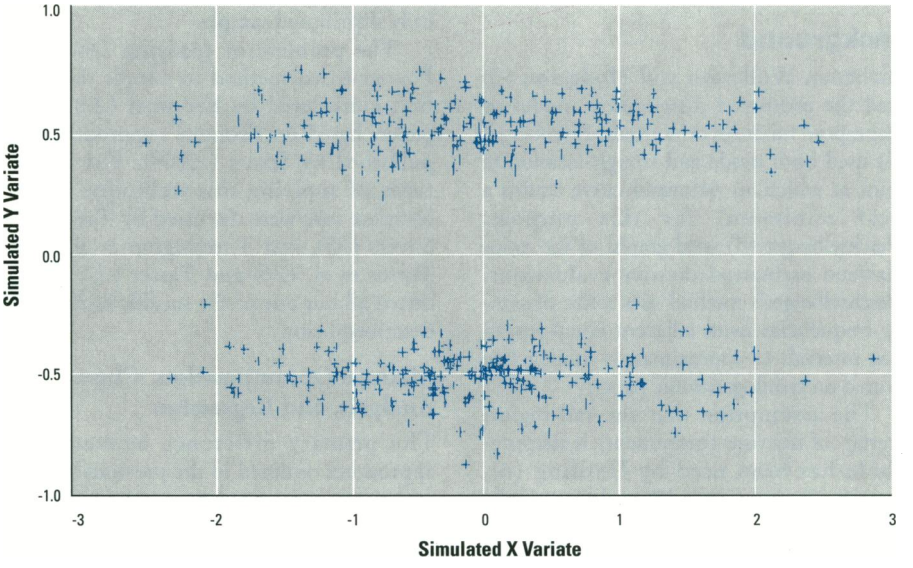

Figure 3. Bivariate data where $P(x) \equiv \Phi(x/\sigma) \rightarrow 0.5$, as $\sigma \rightarrow \infty$, i.e., constant switching.

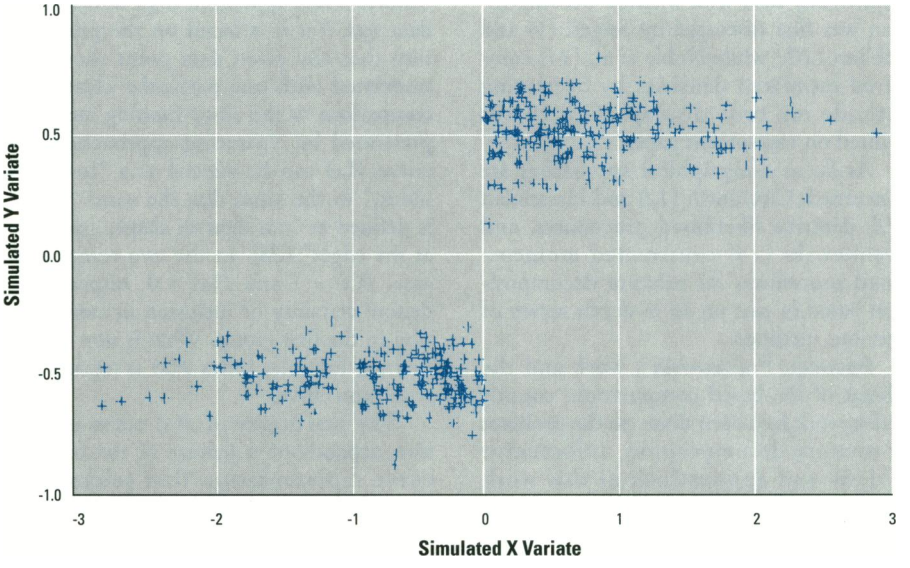

Figure 4. Bivariate data where $P(x) \equiv \Phi(x/\sigma)$, as $\sigma \rightarrow 0$, i.e., sudden switching.

the main, attributable to the upper component. Thus, here, the switching function $P(x)$ is binary valued.

When Φ is chosen to represent the standard normal cumulative distribution function, then Figures 3 and 4 correspond to two limiting cases, where the curve $P(x) \equiv \Phi(x/\sigma)$, and $\sigma \to \infty$, for Figure 3, and $\sigma \to 0$, for Figure 4. On the other hand, the choice of switching function, $P(x) \equiv \Phi(x/1.5)$, was used to obtain the data displayed in Figure 5.

The data shown in Figure 5 may provide a good illustration of how data amenable to conditional switching regression procedures can be simulated, but the estimation of $P(x)$ in this case would be like shooting fish in a barrel. It is an unrealistically easy problem because of the high degree of separation between the two components that was induced by the choice of the two Y-variate standard deviations to be 0.1. Consequently, to provide both a more realistic and a more challenging example, the data shown in Figure 6 were generated.

To obtain Figure 6, the four X and Y-variate standard deviations were all selected to equal 1, the two X means were set equal to 0, while again $P(x) \equiv \Phi(x/1.5)$. It was from this sample that the curve shown as a solid line in Figure 2 was estimated. It is noteworthy that this configuration of data could easily be mistaken for single-component bivariate normal data where there is a positive correlation between variates. Here, quite to the contrary, the correlation between the X and Y variates within each mixture component was set equal to 0. Thus, the pronounced covariance between the two variates is entirely due to the switching regression line $P(x) \equiv \Phi(x/1.5)$.

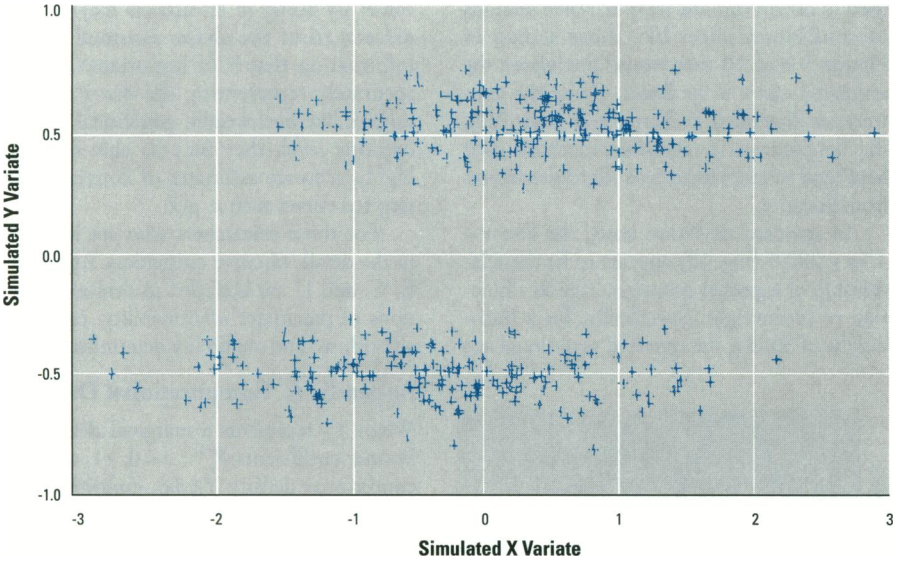

Figure 5. Bivariate example where, since $P(x) \equiv \Phi(x/1.5)$, switching is gradual.

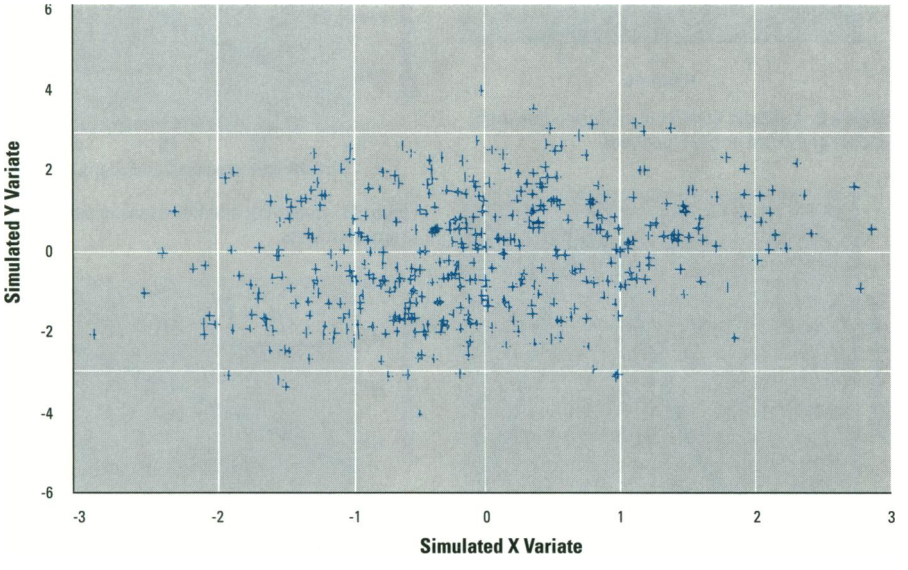

Figure 6. Example of bivariate data where $P(x)$ switches through a continuous sequence of values and the overlap between components is appreciable.

Procedure for the Estimation of $P(x)$

The first step we take to estimate a mixing parameter regression curve is to use procedures described in detail in Tarter and Lock (*17*) to estimate the joint density of the dependent and independent variate. Then Equations 2.22 and 2.23 of Tarter (*29*) are used to estimate the conditionals at a sequence of X variate values (see Appendix A). The slice taken at point $x = 0$ through the line shown in Figure 6 is shown in Figure 7.

Figure 8 illustrates the Kronmal (*22*) and Titterington et al. (*15*) procedure applied to the curve shown in Figure 3. This involved the sample-size-controlled modification of the Fourier transform of the estimated density to obtain a curve estimate where the standard deviations of all mixture components are all reduced by a user-selected constant $\lambda = 0.45$. As the sample size n increases, and for any mixture whose component standard deviations are all greater than this value of λ, these sample standard deviations will all be reduced by 0.45, while all other distributional characteristics will remain the same.

Once the mixture-component-specific variances are reduced sufficiently to assure that the resulting curve has no component overlap, one or another of the components

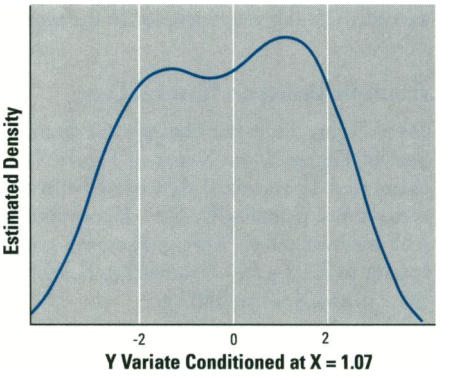

Figure 7. Example of estimated conditional slice.

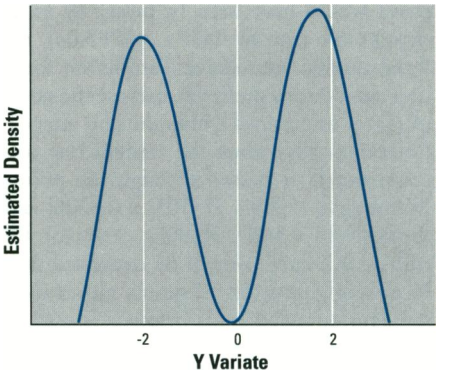

Figure 8. Mixture-component-reduced conditional slice.

can be excised and mixing parameter $P(x)$ can be estimated for the slice at $X = x$. The results of applying this process, first to the leftmost and then to the rightmost component, as well as the fit of normal densities, are shown in Figures 9 and 10. As shown in Figure 2, when evaluated at $x = 0$ the estimated $P(x)$ curve passes through the point $y = 0.5483$. Thus, in this example, while the $P(x)$ curve's slope has been estimated very accurately, and there is almost perfect fit to the appropriate normal density, there is some variation between the estimated y intercept and the true value $y = 0.5$.

The trials described below were based on a sample of 447 northern California Kaiser Permanente-insured, nonsmoking mothers who gave birth from 1960 to 1963. These data were obtained from "Child Health and Development Studies" (*33*) conducted in northern California. The investigation that used the new form of regression was a preliminary to research concerning the relationship between measured blood levels of the biomarker cotinine that might have been induced by ambient cigarette smoke and birthweight adjusted for gestational age. It was postulated that either guessing or some other factor predisposed to division into two subgroups on the basis of stated gestational age. For this reason, gestational age was assigned as a response variate, while birthweight was selected as the key independent variate. Only those births where gestational age exceeded 260 days and birthweight exceeded 2500 g were included in the sample of 447.

Example Based on Natural Data

Since Figure 11 is the natural data counterpart of Figure 2, the X-axis of Figure 11 is calibrated in standard deviation from the mean units. Specifically, X = -1 corresponds to a birthweight of 3004 g, X = -0.5 corresponds to 3207 g, X = 0 to 3407 g, X = 0.5 to 3607 g, and X = 1 to 3807 g.

The $P(x)$ curve was estimated to be an almost perfectly straight line. The best-fitting normal cumulative was found to be $P(x) \equiv \Phi[(x + 3.67)/3.03]$. Although the empirical curve would have been fit better by some straight line than by $\Phi[(x + 3.67)/3.03]$, the fitted normal cumulative distribution function was deemed preferable because the range of $P(x)$ is constrained within the unit interval. Parenthetically, when the straight line was constructed, it passed through the points {birthweight = 3003, $P(3003) = 0.8206$} and {birthweight = 3809, $P(3809) = 0.9446$}. Of course, this curve should be converted into birthweight in units of grams rather than standard deviation units before it is used to predict group membership for a given birthweight. In either scale, however, there is a pronounced decline in upper-group membership with increased birthweight. Sufficient trials with simulated data have been conducted to suggest that the $P(x)$ curve shown in both Figures 11 and 12 does indeed have a positive slope. Unfortunately, as reviewed in Tarter and Lock (*17*), despite the many advances made in the last 10 years in curve-based inferential procedures, confidence band or other conventional ways of assessing the credibility of $P(x)$ are currently outside the realm of practicality.

If upper-group membership is attributable to guessing, then women who guess rather than accurately report gestational age have a marked tendency to also give birth to infants with low birthweights.

Within the putative nonguess subpopulation, the estimated relationship between expected gestational age, $\mu(x)$, and birthweight, shown in Figure 11, was found to be almost exactly linear. There is absolutely nothing about the nonparametric procedures used to obtain this line from repeated sections of conditional slices like those shown in Figures 9 and 10 that would predispose the estimated curve to be linear. Quite the contrary, as illustrated in Tarter and Lock (*17*), the nonparametric regression methods used here tend to emphasize even slight departures from linearity.

In standard deviation units, the line $y = 4.5x + 280.0$ (Fig. 12) appears to fit the relationship of expected gestational age as a function of birthweight. Specifically, for a birthweight of 3003 g the expected gestational age was estimated to be 275.4 days, birthweight 3205 g to be age 277.8 days, 3406 g to be 280.24 days, 3607 g to be 282.3 days and 3809 g to be 284.2 days.

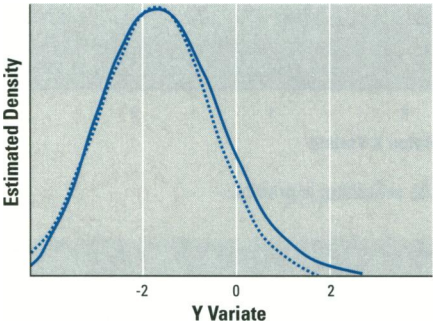

Figure 9. Leftmost estimated mixture component (solid) and fitted normal (dashed).

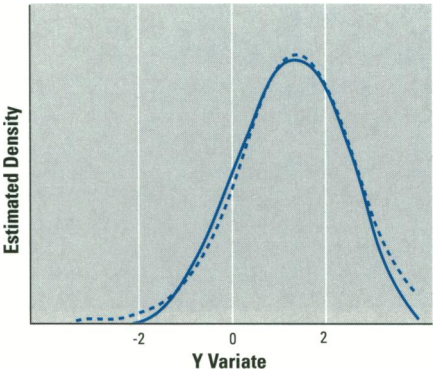

Figure 10. Rightmost estimated mixture component (solid) and fitted normal (dashed).

Discussion

The above example raises a question whose answer has considerable bearing on many environmental studies. This question is best formulated in the context of the data many economists now study with the help of switching regression methodology. Why are the data forms studied by environmental scientists universally dealt with by ordinary single-valued regression curves, when many-valued curves are thought necessary for the study of economic data? Once it is conceded that clusters, subgroups, and their finite mixture representation are relevant, why not go one step further and consider the possibility that the curve $P(x)$ does not assume a constant value? By doing so researchers may not only discern from the shape assumed by $P(x)$ information that is of importance, but, by accurately representing the distribution of variates like birthweight, gestational age, and cotinine level, they may be able to obtain highly accurate estimates of conventionally targeted curves such as $\mu(x)$.

For those researchers who are interested in the details of curve estimation, Appendices B, C, and D are included to deal with questions of parameter identifiability, estimation algorithms, and credibility determination.

Appendix A. Computational Details

When f_m represents a marginal density, the Fourier coefficients $B_v^{(x_0)}$; $v = 0, \pm1, \pm2...$; of a conditional density $f_c(y|x)$, conditioned on

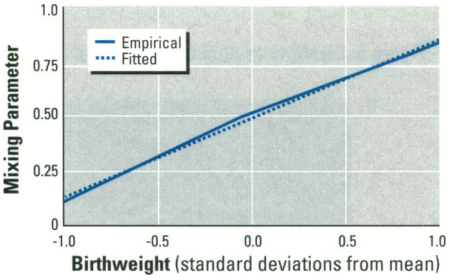

Figure 11. Empirical and fitted mixing parameter regression lines.

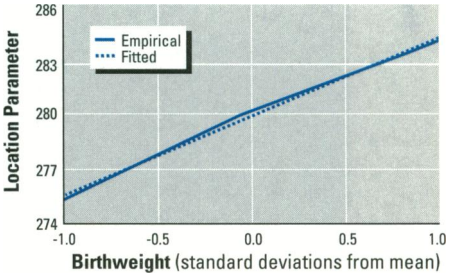

Figure 12. Empirical location parameter and fitted $y=4.5x + 280.0$ regression.

the key variate value $X = x_0$, have a simple relationship to the coefficients B_{uv}; $u = 0, \pm1, \pm2...$; $v = 0, \pm1, \pm2...$, of the joint density $f(x,y) = f_c(y|x)\, f_m(x)$. Assuming that $f_c(y|x_0)$ can be represented as

$$\sum_{v=-\infty}^{\infty} B_v^{(x_0)} e^{2\pi i v y},$$

suppose the coefficients, B_{uv} $u = 0, \pm1, \pm2...$; $v = 0, \pm1, \pm2...$, can be substituted into the formula

$$B_v^{(x_0)} = \frac{1}{f_m(x_0)}\left[\sum_{s=-\infty}^{\infty} B_{s,v}\, e^{2\pi i s x_0}\right] \quad \text{(A.1)}$$

At the value $x = x_0$, estimates of $B_v^{(x_0)}$; $v = 0,\pm1, \pm2...$; can be used to determine mixing parameter $P(x_0)$ as follows:

1) The product is formed of each $B_v^{(x_0)}$ times a multiplier that solely affects the overlap of mixture component variances. The history of this lambda multiplier system is outlined in the Background section.

2) Series stopping and term inclusion rules based on the resulting product are applied.

3) The smallest positive user control setting λ of the lambda multiplier system that reduces component overlap to zero at some value of y between the two-mixture component mean values is determined. (There is little chance that this optimal value of λ is multiple-valued rather than single-valued.)

4) The component variance-reduced density determined in steps 1), 2), and 3) is multiplied by the indicator function, $I_{[a,b]}(x) = 1$, for $a \le x \le b$, 0 elsewhere; where the interval $[a,b]$ is the support region of one of the two components. This step blanks out one of the mixture components and isolates the other.

5) The control setting $-\lambda$ of the lambda multiplier system is used to snap the isolated component back into place.

6) The function constructed by means of executing steps 1–5 is a Fourier series. Consequently, the integral of this curve over its support region can be determined without the need for numerical integration techniques. This integral provides the estimate of $P(x_0)$.

Detailed procedures for performing these 6 steps are described by Tarter and Lock (*17*), as are certain theoretical issues involved. Specifically, step 1 is discussed, which deals with the issue of multiplier system independence. As far as the crucial step 2 is concerned, a technique for estimating the variances of estimators of conditional coefficients $B_v^{(x_0)}$; $v = 0,\pm1, \pm2...$; is outlined (*17*). Knowledge of these variances is required to determine stopping and term inclusion rules.

Appendix B. Assumptions and Parameter Identifiability

The question of mixing parameter identifiability is related to the distinction between curve properties, such as the variances of mixture components, and the parameters that form part of mixture representation. For the c-component mixture model

$$\begin{aligned} f\left(y|p_1, p_2, ..., p_{c-1}, \mu_1, \mu_2, ..., \mu_c, \sigma_1, \sigma_2, ..., \sigma_c\right) &= p_1 f_1\left[\left(y-\mu_1\right)/\sigma_1\right]/\sigma_1 \\ &+ p_2 f_2\left[\left(y-\mu_2\right)/\sigma_2\right]/\sigma_2 \\ &+ ... + \left(1 - \sum_{j=1}^{c-1} p_j\right) f_c\left[\left(y-\mu_c\right)/\sigma_c\right]/\sigma \end{aligned} \quad \text{(B.1)}$$

where $c > 1$, the validity of the decomposition procedure depends on two classes of assumptions. The first is that unimodal f_1, f_2,... are symmetric about a mode with coordinate zero. The second is that the parameters σ_1^2, σ_2^2 ,..., σ_c^2, are individually proportional to, or have some other monotonic relationship to, the population variances (second cumulants) of the respective subcomponents. The second assumption concerns a relationship between 1) the parameters of f_1 through f_c's parametric representations and 2) the specific properties, the variances, of the curves f_1 through the f_c.

In particular, assume that for $s = 1,...c$ the sth characteristic function of $f_s(x)$ can be expressed as $\exp[w_s(iz)]$, where the jth cumulant κ_{js} of the sth component is assumed to exist and be well defined. (The sth cumulant of a density f is the evaluation at zero of the sth derivative of the log of the moment-generating function of f.) Then by the definition of cumulants given by Kendall (*34*)

$$W_s(z) = \sum_{s=1}^{c} \kappa_{js} z^j / j! \quad \text{(B.2)}$$

For example, if $f_s(x) = \Phi[(x-\mu_s)/\sigma_s]/\sigma_s$, $\Phi(x) = N(x|0,1)$, in other words, the standard normal; then $W_s(z) = \mu_s z + \sigma_s z^2/2$.

Combining the above relationships, suppose a certain sample of iid data with datum density f has been transformed to the unit interval in such a way that for all practical purposes the kth Fourier coefficient of the expansion of f equals

$$B_k = \sum_{s=1}^{c} p_s \exp\left[-2\pi i k \mu_s - 2\pi^2 k^2 \sigma_s^2 - \sum_{j=3}^{\infty} \kappa_{js} (ik)^j / j!\right] \quad \text{(B.3)}$$

For the sequence of multipliers

$$\left\{b_k^{(\lambda)} = \exp\left[2\pi^2 k^2 \lambda^2\right]\right\} \quad \text{(B.4)}$$

when f_λ is defined to be the function whose Fourier expansion has coefficients $\{b_k^{(\lambda)} B_k\}$, it follows that each member of this sequence of coefficients equals

$$\sum_{s=1}^{c} p_s \exp\left[-2\pi i k \mu_s - 2\pi^2 k^2 \left(\sigma_s^2 - \lambda^2\right) - \sum_{j=3}^{\infty} \kappa_{js} (ik)^j / j!\right] \quad \text{(B.5)}$$

To interpret the effect of the multiplier sequence (Eq. B.4) on the function whose kth Fourier coefficient is B_k, notice that the one difference between Equations B.3 and B.5 is that in Equation B.5 the term $(\sigma_s^2 - \lambda^2)$ appears instead of the term σ_s^2.

The roles played by multiplier sequence (Eq. B.4) can be explained by noting that the term $(\sigma_s^2 - \lambda^2)$ of Equation B.5 is associated with an important distributional property of the underlying mixture model. With one exception $(\sigma_s^2 - \lambda^2)$ is the second cumulant, variance, of the sth component of a mixture identical to f. This exception is that the original second cumulants of each original component have each been reduced by the quantity λ^2. Extensions of this procedure to mixtures where components do not have a finite variance are discussed in Tarter and Lock (*17*).

Appendix C. Estimation Algorithms in Terms of Kernels

Figure C.1 was taken from the Tarter and Lock monograph (*17*) and contains graphs of several kernels. These kernels were formed by truncating the Fourier expansion with coefficients $\{b_k^{(\lambda)} = \exp[2\pi^2 k^2 \lambda^2]\}$; $k = 0, \pm1, \pm2.., \pm m$; at either $m = 4$ or $m = 8$. In general, the procedure described above can be used in conjunction with any curve estimator $\hat{f}$ by taking the Fourier expansion of $\hat{f}$ and multiplying each consecutive term of this expansion by the corresponding member of the sequence $\{b_k^{(\lambda)} = \exp[2\pi^2 k^2 \lambda^2]\}$.

In the univariate case, analogs to the kernels shown in Figure C.1 can be obtained by dividing the products described in the previous paragraph by the sample trigonometric moments. If the multipliers formed in this way have real and imaginary components that are both less than one in absolute value, then these components can be treated as the Fourier coefficients of a kernel whose role, like that of all kernels, is to spread the effect of each individual observation over the axis used to represent the variate's density. For those researchers who are familiar with the processes of kernel-based curve estimation, the above procedure can provide the means to adapt curve-estimation software to construct decomposition algorithms.

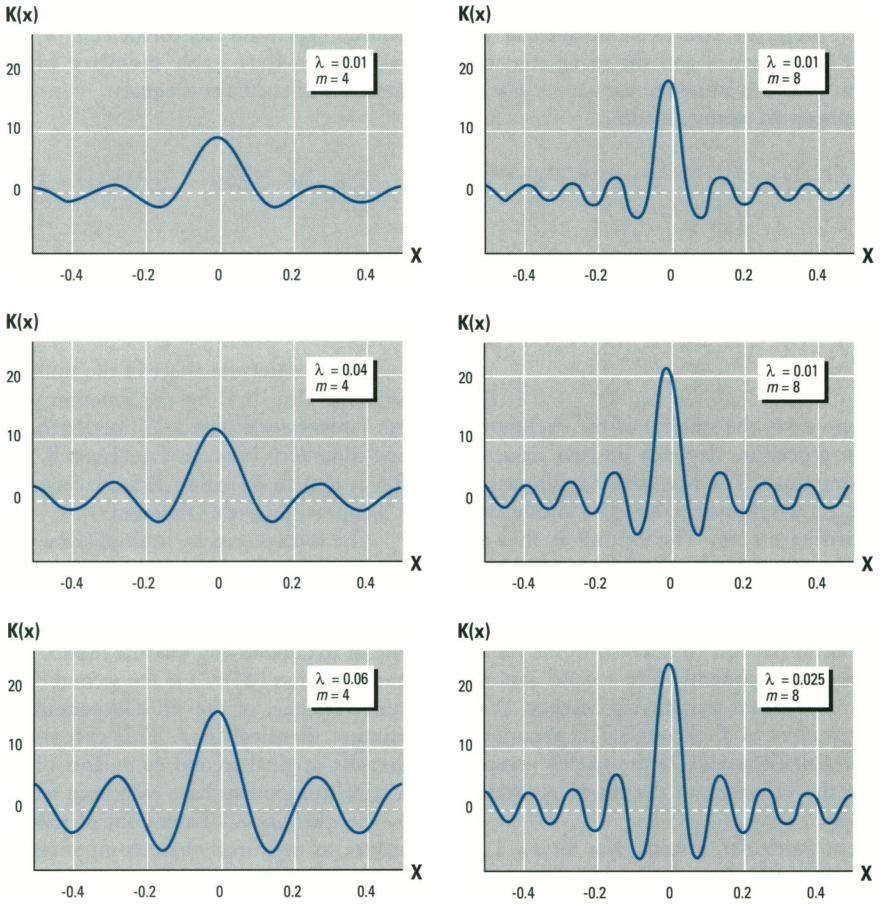

Figure C-1. Kernels used by the mixture decomposition algorithm.

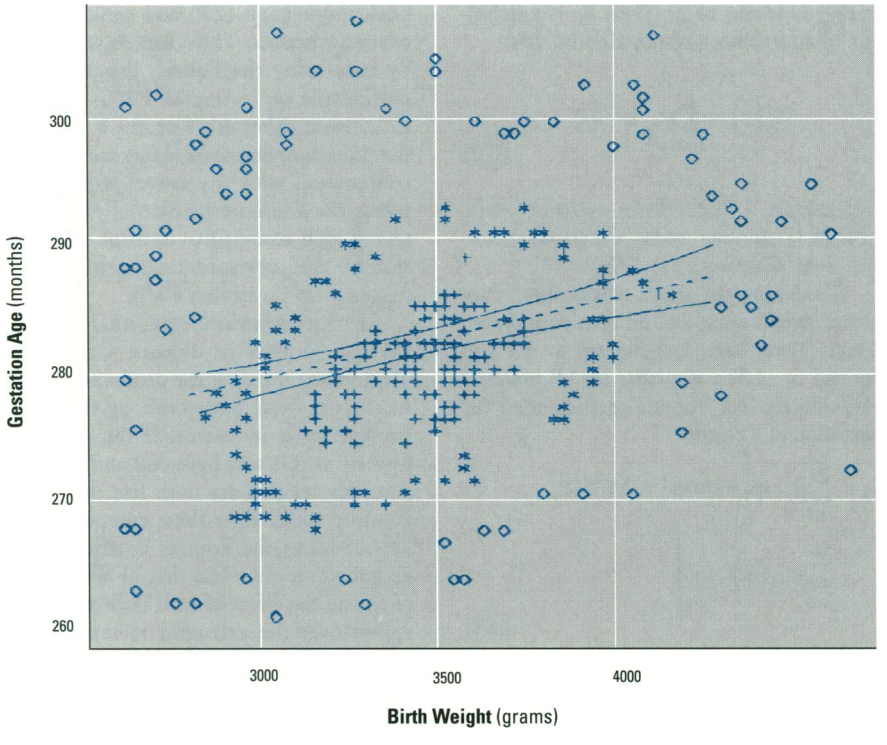

Figure D-1. Stick-pin plot of birthweight and gestation age data with conventional least-squares line and brands.

Appendix D. Credibility

Tarter and Marshall (*35*) describe a general approach that can be used to bootstrap any curve-estimation procedure that is based on Fourier-series methodology. Although this approach has been implemented on a large mainframe computer system, it is currently impractical to use it in conjunction with a PC system. However, by means of a simple non-bootstrap procedure, curves that are related to mixing-parameter forms of regression can be generalized to check on one particular form of credibility. The following is a brief introduction to this procedure in the context of the application to birthweight and gestation age.

Before turning to the topic of credibility assessment, some discussion of a key feature of Figure D.1 is in order. This feature allows distributional contours to be portrayed by means of the data points themselves. In Figure D.1 the position of the displayed points and the plotting character (for example, plus sign, asterisk, or small diamond) used to represent each point is affected in two ways by the estimator of the joint density $f(x,y)$ described in Appendix A.

Assume that the expression $\hat{f}_{(Mode)}$ represents the largest value assumed by the estimator of f, $\hat{f}$, over the unique sample mode. To begin the process of determining plotting characters, data points are substituted as the arguments of $\hat{f}$. Plotting character choices were made in Figure D.1 by subdividing the interval that begins with 0 and ends with $\hat{f}_{(Mode)}$ into five equal-length, contiguous, nonintersecting subintervals and then using these subintervals as follows: a particular (X,Y) point substituted into $\hat{f}$ yields a value that falls within one and only one of the above five intervals. When this value includes or is between 0.8 $\hat{f}_{(Mode)}$ and $\hat{f}_{(Mode)}$ itself, the plotting character chosen to represent this point is a plus sign. However, when the evaluation of $\hat{f}$ yields a value between 0.6 $\hat{f}_{(Mode)}$ and 0.8 $\hat{f}_{(Mode)}$, and yet does not equal either of these values exactly, a null plotting character is used. In other words, the band of points at this intermediate level of estimated density is blanked out.

The reasoning that underlies this "blank banding" process is that this ring of points is less informative than is the border that the eye sees between the blanked region and the unblanked region above and, as will now be defined, the border below the unblanked region. When the estimated density is between 0.4 $\hat{f}_{(Mode)}$ and 0.6 $\hat{f}_{(Mode)}$, an asterisk is used as a plotting character, and when the estimated density includes and is between 0 and 0.2 $\hat{f}_{(Mode)}$, a small diamond is used. All other points are blanked out.

The blank-banding procedure was originally proposed in Tarter and Freeman (*36*) as a computational trick to produce bivariate contours on simple personal computer sys-

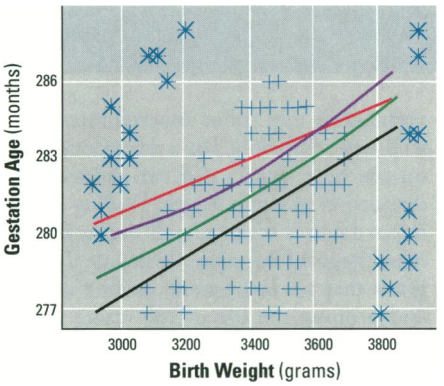

Figure D-2. Zoomed birthweight and gestation age-adapted bands, least squares and nonparametric regression curves. Purple line: upper nonparametric band; red line: least squares regression line (same as purple line); green line, nonparametric regression line; black line, lower nonparametric band.

tems. For the birthweight-gestation age study, these contours help compare conventional 95% standard confidence bands

$$\bar{Y}_x \pm t_{1-\alpha/2} s_{y|x} \sqrt{\frac{1}{n} + \frac{\left(X - \bar{X}\right)^2}{SSX}} \qquad \text{(D.1)}$$

around the simple unweighted least-squares linear regression curve (*37*), to global distribution features. Within the confidence band formula, $\bar{Y}_x$ represents the regression curve that estimates how the gestation Y-variate expectation changes with X, $t_{1-\alpha/2}$ represents the $1 - \alpha/2$ quantile of the Student's t-statistic with $n - 2$ degrees of freedom, $\bar{X}$ is the sample mean, here of the birthweight X variate, and SSX is the estimated X-variate standard deviation.

The apparent translation of the regression line and bands above the centroid of the cluster of points marked with the + plotting character is attributable to the attraction of the guess subgroup or cluster. There are many aspects of estimator credibility that apply to conditional switching regression models. This is because besides the mixing parameter regression curve $P(x)$, at least four other curves, two for each location parameter μ and two for each scale parameter σ, apply to two component mixtures.

Of these four curves, the curve $\sigma(x) = \sigma_x$, which characterizes the scale change with x of the nonguess subpopulation, is the curve that is most closely related to credibility assessment. This is because $\sigma(x) = \sigma_x$ is the curve that is estimated by the crucial component $s_{y|x}$ within the standard confidence band formula (Eq. D.1).

To deal with the $\sigma(x) = \sigma_x$ curve nonparametrically, the term $s_{y|x}$ can be estimated in much the same way that $P(x)$ is estimated and treated as a function of x and not, as it is in conventional regression applications, as a constant. In this way the width of the bands shown about the nonparametric regression curve shown in Figure D2 permits heteroscedasticity to be examined graphically. In the case of the birthweight-gestation age relationship, the width of the confidence bands about the nonparametric regression curve remains the same as that based on (Eq. D.1), even though, as will be described in a future paper, the nonparametric curve is less attracted by the guess cluster than is the least-squares line.

References

1. Tarter ME, Cooper RC, Freeman WR. A graphical analysis of the interrelationships among waterborne asbestos, digestive system cancer and population density. Environ Health Perspect 53:79–89 (1983).
2. Tarter ME. Comment on density estimation and bump-hunting by penalized likelihood method exemplified by scattering and meteorite data by Good and Gaskins. J Am Stat Assoc 75:63–65 (1980).
3. Buchanan-Wollaston HG, Hodgeson WC. A new method for treating frequency curves in fishery statistics, with some results. J Conserva-tion 4:207–225 (1929).
4. Eisenberger I. Genesis of bimodal distributions. Technometrics 6:357–363 (1964).
5. Weichselberger VK. Uber ein graphisches verfahren zur trunnung von mischvertielungen und zur identifikation kupietier normal-verteilungen bei grossem stichprobennumfang. Metrica 3:178–229 (1961).
6. Harding JP. The use of probability paper for graphical analysis of polymodal frequency distributions. J Mar Biol Assoc 28:141–153 (1949).
7. Preston EJ. A graphical method for the analysis of statistical distributions into two normal components. Biometrika 40:460–464 (1953).
8. Cassie RM. Some uses of probability paper in the analysis of size frequency distributions. Aust J Mar Freshwater Res 5:513–522 (1954).
9. Szigeti G. Lumeniszkalo anayagod. Elekrok-technika 39:70–73 (1947).
10. Labhart H. Ein auswetegerat fur elektrophoresdiagramme. Experientia 3:36–37 (1947).
11. Noble FW, Hayes JE, Eden M. Repetitive analog computer for the analysis of sums of distribution functions. Proc Inst Radio Eng 47:1952–1956 (1959).
12. Cicchinelli AL. The composite of two gaussian distributions as a model for blood pressure distributions in man (PhD dissertation). Lansing, MI:University of Michigan, 1962.
13. Hasselblad V. Estimation of parameters for a mixture of normal distributions. Technometrics 8:431–444 (1966).
14. Frankowski R. The estimation of parameters in a mixture of gaussian distributions with applications to the distribution of serum uric acid in man (PhD dissertation). Ann Arbor, MI:University of Michigan, 1967.
15. Titterington DM, Smith AFM, Makov UE. Statistical analysis of finite mixture distributions. Chichester, England:Wiley, 1985.
16. Tukey JW. A problem of Berkson, and minimum variance orderly estimators. Ann Math Stat 29:588–592 (1958).
17. Tarter ME, Lock MD. Model-free curve estimation. New York:Chapman and Hall, 1993.
18. Medgyessy P. Decomposition of superpositions of distribution functions. Budapest:Publishing House of the Hungarian Academy of Sciences, 1961.
19. Doetsch G. Die elimination des dopplereffekts auf spektroskopische feinstrukturen und exakte bestimmung der komponenten. Zeitschrift Phys 49:705–730 (1928).
20. Doetsch G. Zerlegung einer function in gauss'sche fehlerkurven. Math Z 41:283–318 (1936).
21. Gregor J. An algorithm for the decomposition of a distribution into gaussian components. Biometrics 25:79–93 (1969).
22. Kronmal RA. The estimation of probability densities (PhD dissertation). Los Angeles, CA: University of California, 1964.
23. Stanat DF. Nonsupervised pattern recognition through the decomposition of probability functions (dissertation no. 7683). Ann Arbor, MI:University of Michigan Sensory Intelligence Laboratory, 1966.
24. Stanat DF. Unsupervised learning of mixtures of probability functions. In: Pattern recognition (Kanal LN, ed). Washington, DC: Thompson Book Company, 1968;357–389.
25. Tarter ME, Silvers A. Implementation of bivariate gaussian mixture decomposition. J Am Stat Assoc 70:47–55 (1975).
26. Tarter ME, Rigsbee EO, Wong JT. Interactive editing of biomedical data. Comput Programs Biomed 6:117–123 (1976).
27. Tarter ME. Biocomputational methodology—an adjunct to theory and applications. Biometrics 35:9–24 (1979).
28. Procter P. Longman dictionary of contemporary English. London:Longman Group Limited, 1978.
29. Tarter ME. Trigonometric maximum likelihood estimation and application to the analysis of incomplete information. J Am Stat Assoc 74: 132–139 (1979).
30. Quandt RE. The estimation of the parameters of a linear regression system obeying two separate regimes. J Am Stat Assoc 51:873–880 (1958).
31. Quandt RE. New approach to estimating switching regressions. J Am Stat Assoc 67:306–310 (1972).
32. Quandt RE, Ramsey JB. Estimating mixtures of normal distributions and switching regressions. J Am Stat Assoc 73:730–751 (1978).
33. Child Health and Development Studies. Data archive and users manual of the child health and development studies, version 1.1. Berkeley, CA:University of California, 1987.
34. Kendall MG. The advanced theory of statistics, vol 1, 5th ed. New York:Hafner Publishing Company, 1952.
35. Tarter ME, Marshall J. A new procedure designed for simulation studies. Commun Stat Ser B 7:283–293 (1978).
36. Tarter ME, Freeman W. On graphing estimated distributions using modified scatter diagrams. Comp Stat Data Anal 7:51–67 (1989).
37. Dixon W, Massey FJ. Introduction to statistical analysis. New York:McGraw-Hill, 1983.